The Aspergillus Mine - publishing bioinformatics.

Vesth, Tammi Camilla; Rasmussen, Jane Lind Nybo; Theobald, Sebastian; Kjærbølling, Inge; Frisvad, Jens Christian; Nielsen, Kristian Fog; Lyhne, Ellen Kirstine; Kogle, Martin Engelhard; Kuo, Alan; Riley, Robert

Total number of authors:
16

Link to article, DOI:
10.7490/f1000research.1113115.1

Publication date:
2016

Document Version
Publisher's PDF, also known as Version of record

The Aspergillus Mine - Publishing bioinformatics

Tammi C. Vesth* (1), Jane L. Nybo (1), Sebastian Theobald (1), Inge Kjærbølling (1), Jens C. Frisvad (1), Kristian F. Nielsen (1), Ellen K. Lyhne (1), Martin E. Kogle (1), Alan Kuo (3), Robert Riley (3), R.P. de Vries (4), Igor V. Grigoriev (3), Uffe H. Mortensen (1), Bernard Henrissat (5), Scott E. Baker (2), Mikael R. Andersen (1)

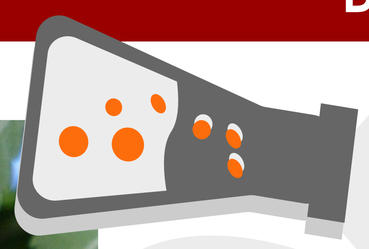
1) Department of Systems Biology, Technical University of Denmark, Kgs. Lyngby, Denmark
2) Joint Bioenergy Institute, Berkeley, CA, USA
3) Joint Genome Institute, Walnut Creek, CA, USA
4) Fungal Physiology, CBS-KNAW Fungal Biodiversity Centre, Utrecht, the Netherlands

Technical University of Denmark · JGI Joint Genome Institute Department of Energy · jbei Joint BioEnergy Institute · CBS-KNAW Fungal Biodiversity Centre, An institute of the Royal Netherlands Academy of Arts and Sciences

Curious about the Aspmine? Ask me!

tcve@bio.dtu.dk

Background



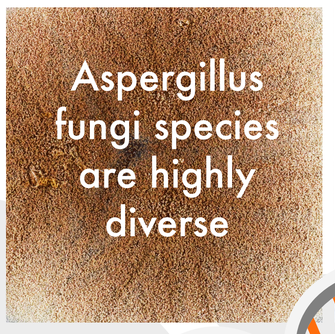
Fungi

Producers of **chemically** and **medically** relevant **compunds**

Aspergilli

Well studied **production organisms**

Can be genetically optimized to produce **cheeper** and environmentally friendly **medicin** and **chemicals**



Sequencing

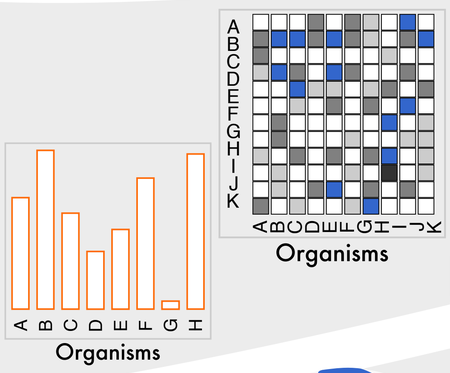
Inititative to sequence **> 300 species**

200 million years of evolution

DTU IBT Culture Collection
> 35,000 fungal cultures
Penicillium, **Aspergillus**, Fusarium, Alternaria and Tricoderma

Comparative Genomics

- Shared genes/functions
- Gene content/genome statistics
- Species specific genes
- Secondary metabolism clusters
- Horizontal gene transfers
- Families of genes and clusters
- Sequence data quality
- Genome sequence diversity



Knowledge sharing

- Cheap hosting of **Interactive** web-applications
  - **Results** and **analysis** with customization
    - Comprehensive **documentation**

Approach - connecting data

Relational database

Strain information
- Raw sequences
- Taxonomy

Functional lables & annotation
- Gene Ontology
- KEGG pathways
- KOG groups
- InterPro domains
- PFAM domains
- SMURF secondary metabolism

BLAST alignment scores
- All against all alignments
- alignment coverage > 50%
- Query/hit coverage
- Percent identity
- Reciprocal hit flag

Q--K--E--S--G--P--S--S--S--Y--C
V--Q--Q--E--S--G--L--V--R--T--T--C

PLATFORM MySQL, Python & R

Analysis

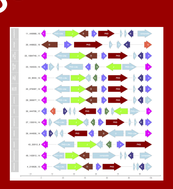
- Conditional data selection
- Data joining
- Comparative measures
- Protein families
- Cluster families

&
- Specific bindingsite
- Specific cellular localization
- Homologs in set of species

&
- Specific function
- Found in species set
- Proximity to other gene

PLATFORM R & RShiny

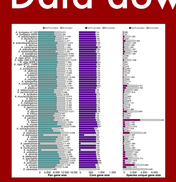
Visuals & Publication



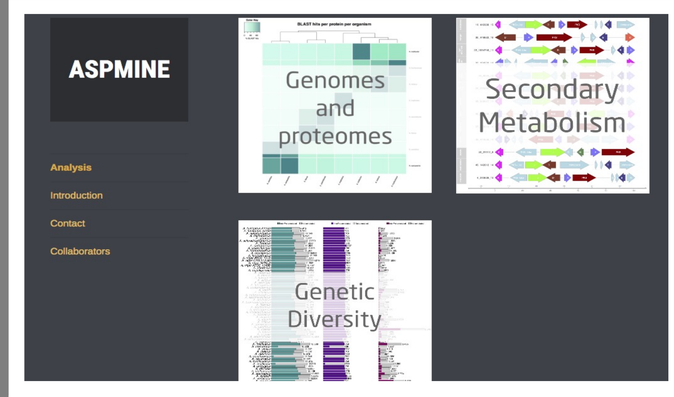
Documentation & analysis

Free online hosting with Wordpres

Data download, figures & interactive figures:



Rshiny hosting with ShinyIO, $440/Year
- Unlimited Applications
- 500 Active Hours
- Performance Boost



Documentation

Data science in biology is a increasing

Publication of thorough data methods is insufficient

Difficulties in reproduction of results

The Aspergillus Mine offers online documentation for each analysis

Online access to analysis data and descriptions of methods

aspmine.wordpress.com

Genetic Diversity

DNA/protein sequence comparisons are essenitial to comparative genomcis

Similar sequences imply similar biological function

Proteins with similar sequences form clusters of functinally related proteins

Core: Families with one or more proteins from each strain
Unique: Families with proteins from only one strain

Closely related strains share more families. Many families are strain specific!

fungiminions.shinyapps.io/geneticDiversity

Families of proteins

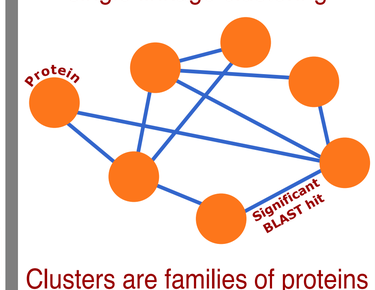
Sequence alignments - BLAST

Q--K--E--S--G--P--S--S--S--Y--C
V--Q--Q--E--S--G--L--V--R--T--T--C

Alignment significance cutoffs

Coverage hit + query > 130% & Percent identity > 50%

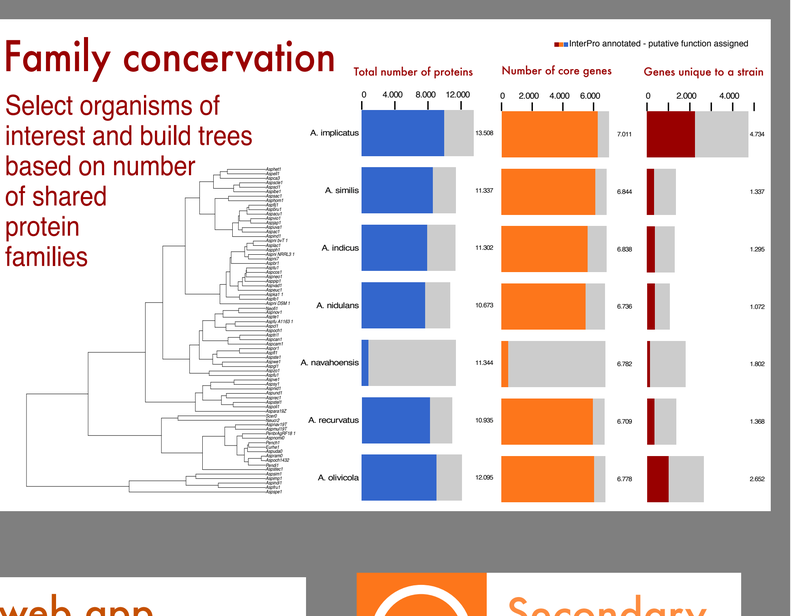
All against all protein alignments and single linkage clustering

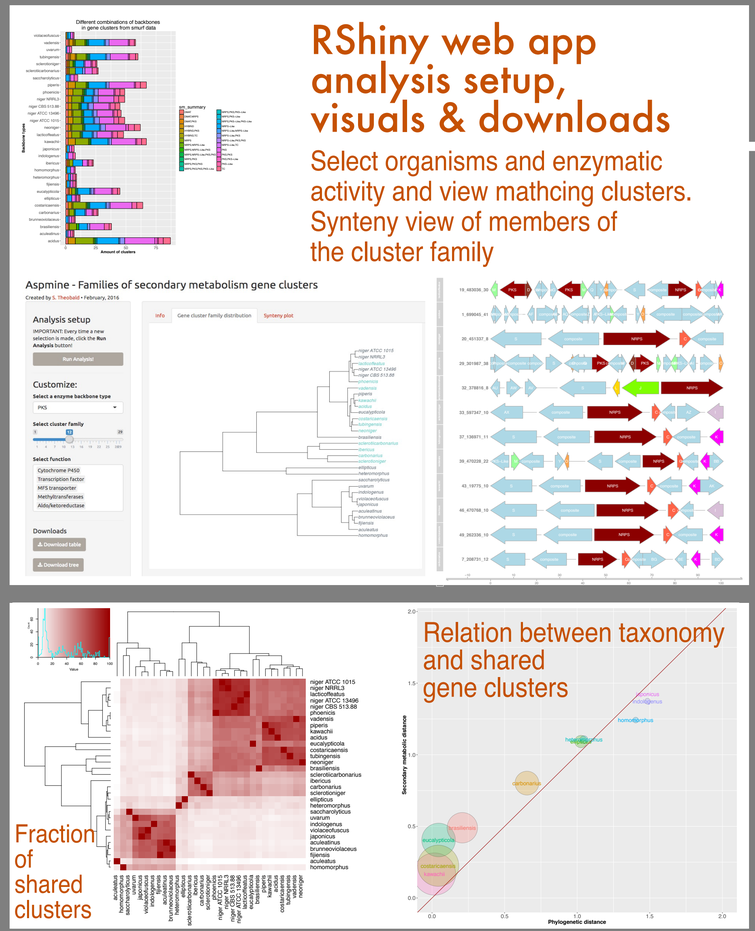


Clusters are families of proteins

Family concervation

Select organisms of interest and build trees based on number of shared protein families



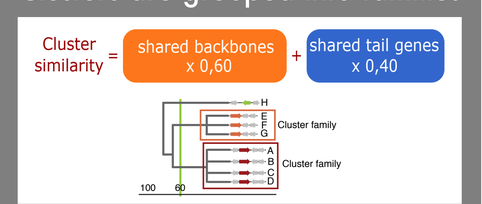


Secondary Metabolism

Secondary metabolites are important medical and chemical compounds

Encoded by gene clusters

Clusters are grouped into families



Gene clusters are highly diverse, do not follow taxonomy and illustrate mechanisms for speciations

fungiminions.shinyapps.io/familiesPhyloTreeApp

aspmine.wordpress.com

www